Comparing Name- and Education-based Country Classification

Henry Chen

April 7, 2023

Contents

1 Introduction:

Classification of individuals based on their country of origin is a crucial aspect of this working paper. There are different two methods tried to determine an individual's country of origin, including using their earliest education degree or analyzing their last name using the NamePrism tool. The former approach assumes that an individual's country of education corresponds to their country of origin, while the latter approach utilizes statistical models to estimate an individual's country of origin based on their last name.

In this task, we aim to compare the performance of these two methodologies and investigate the correlation between them. We will present a summary table and a correlation matrix to provide insights into the effectiveness of these methods in classifying individuals based on their country of origin. Such an analysis is valuable in evaluating the validity of these methods and identifying potential biases that might affect their accuracy.

2 Task:

The task is laid out as follows:

- Bridging the country of origin in the founder dataset and geo-regions defined in NamePrism
- Calculating the matching score from two classification methods
- Show those unmatched cases in detail (e.g. distribution)
- Draw some examples where NamePrism says are English but educated elsewhere
- Draw some examples but the opposite: Education says English but NamePrism elsewhere
- Calculate what share of education-based countries match the 1st vs. the 2nd vs. the 3rd highest-probability name-based country
- Calculate the share of match conditional on NamePrism Matching Probability Threshold (25%, 50%, and 75%)

Files to do crossmatch

- Education-based classification: **country_classification_education**
- Last Name-based classification: **lastname_nationality**

3 Results:

3.1 Files for Analysis:

The comparison process involves the use of two files, namely, the code and the bridge. The code is a Stata do file that has been designed to perform comparisons and generate graphs for distributions and heatmaps. On the other hand, the bridge is a bridging file that has been manually created to establish a connection between the Nationality Taxonomy in Name Prism and the Countries of origin listed in founder data.

To use these two files effectively, you will need to open the code and include your workspace directory. Once this is done, you can proceed to run the code to obtain the desired results and graphs.

3.2 Comparison Results Summary

First of all, Figure 1 and 2 show the Nationality Taxonomy distribution under two classification methods. To be noted, Figure 1 has a large spike in the "International" category. There is a large share in this specific category because of two reasons:

- 1) 78% of the founder's origin country is classified as "United States";
- 2) "United States" is included in "International" in Nationality Taxonomy.

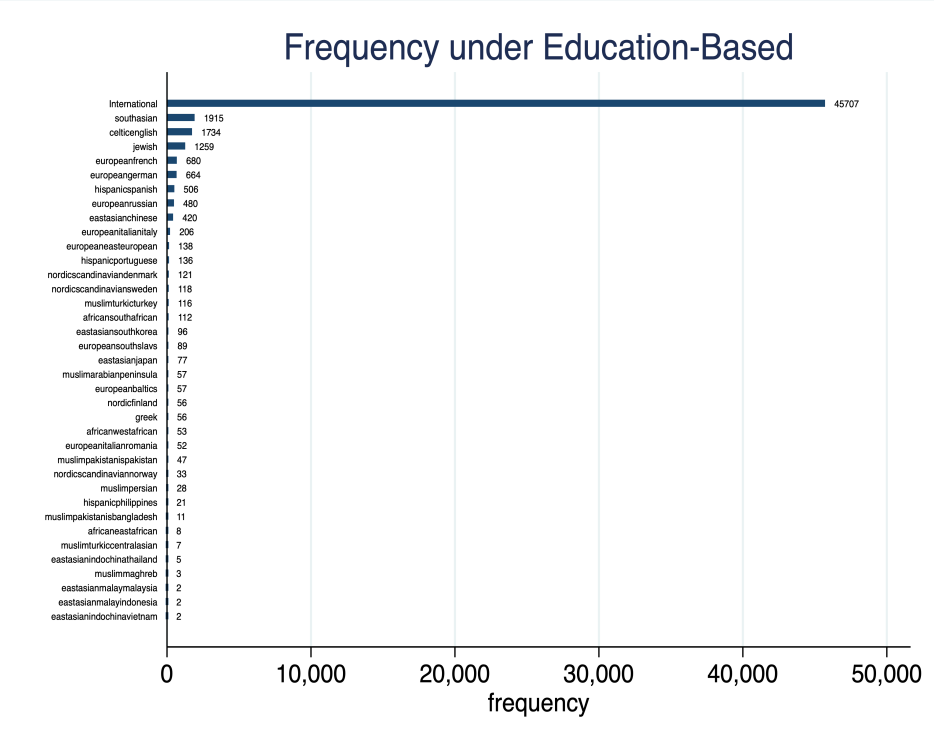

Figure 1: Distribution under Education Classification

Therefore, the education-based classification assigns a large portion of the founder to "International" as their country of origin. However, it should be noted that the "International" category encompasses more countries than just the "United States." As shown in Figure 3, the education-based classification includes 19 different countries in the "International" category. By excluding observations in the "International" category under the education-based classification, it appears that the two classification methods generate similar distributions overall, as depicted in Figure 4. I conducted a matching analysis between the name-based and education-based classifications and found that **55%** of the sample (9367 observations when excluding "International," and 55074 for all) was a match. Figure 5 displays how these matches are distributed across the Nationality Taxonomy categories.

Turning attention to the unmatched observations, Figure 6 presents the distribution under each classification separately. To gain a better understanding of how the two classification methods misalign, Figure 7 displays a heatmap of all combinations of classified taxonomies under both methodologies. The Y-axis represents each category assigned under the education-based method, while the X-axis describes the categories assigned under the name-based algorithm.

Figure 7 reports some patterns worthy noticed.

- Among those unmatched, the name-based classification tends to classify founders into three segments:

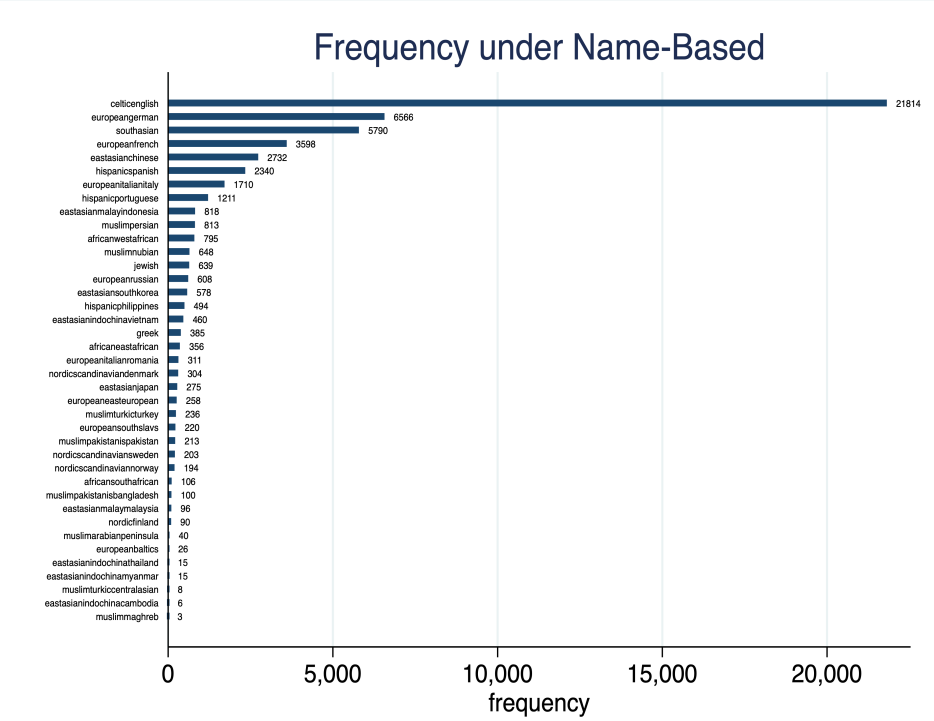

Figure 2: Distribution under Last Name Classification

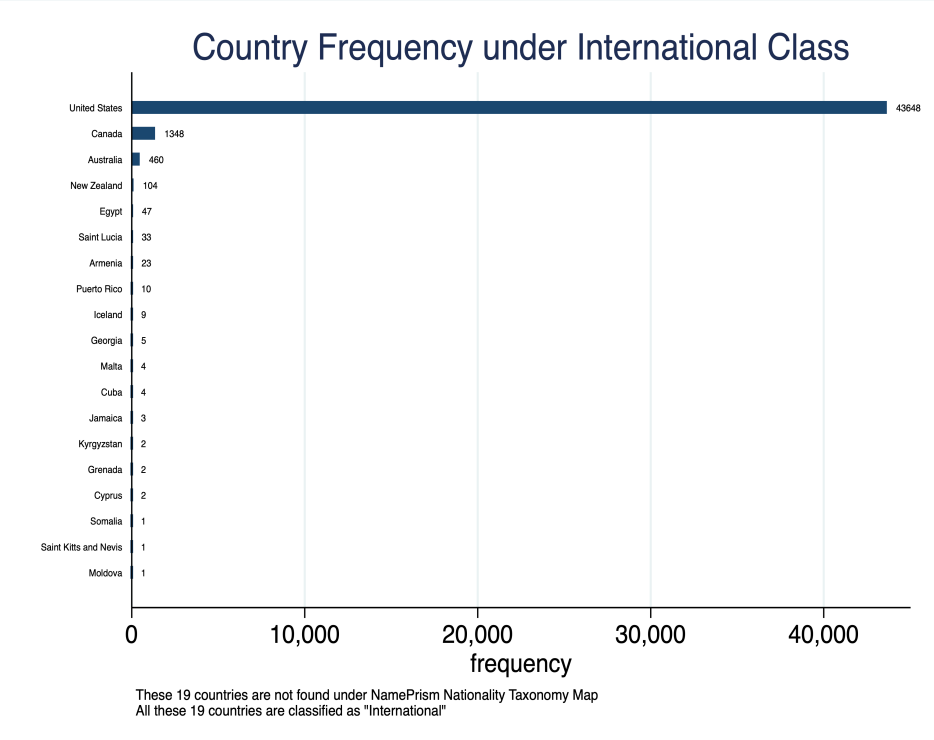

Figure 3: Distribution within International Category

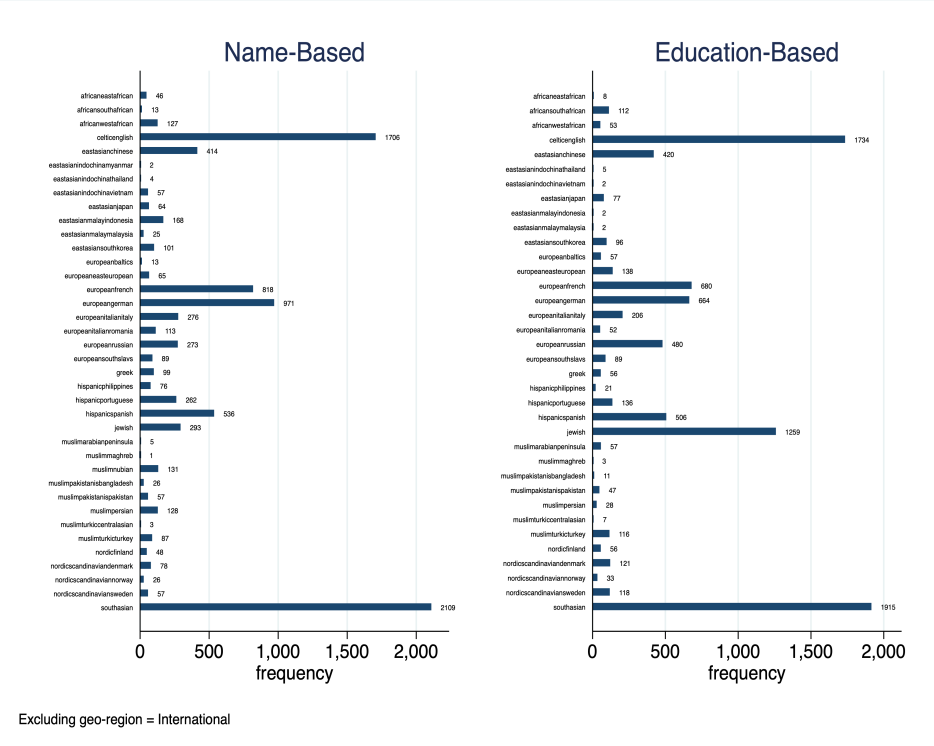

Figure 4: Distribution excluding International Category

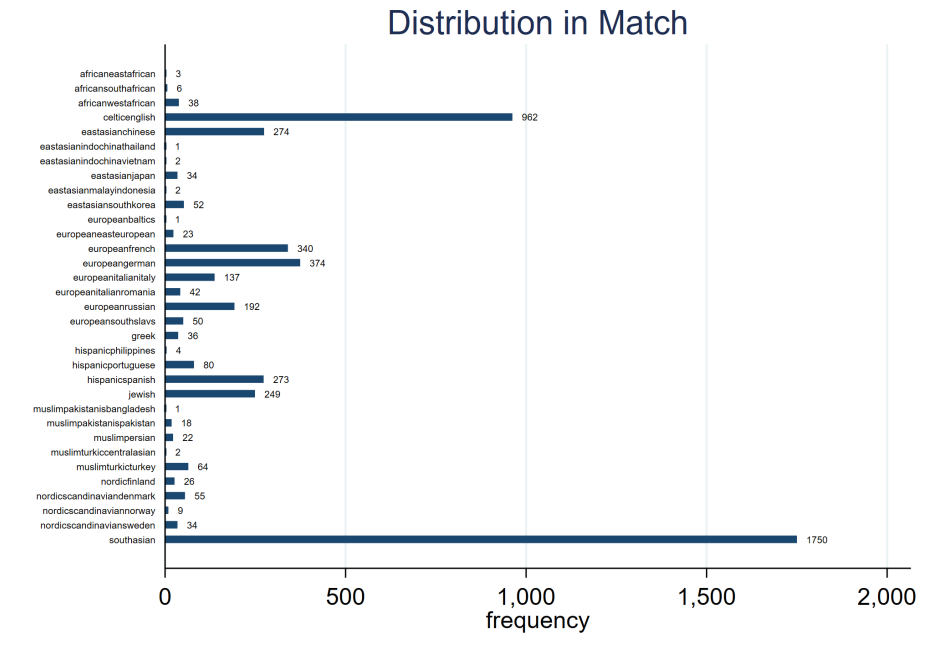

Figure 5: Distribution for Match

  - Celtic English
  - European German
  - European French
- The education-based one, on the other hand, tends to have misaligned assignments in
  - Jewish
  - Celtic English

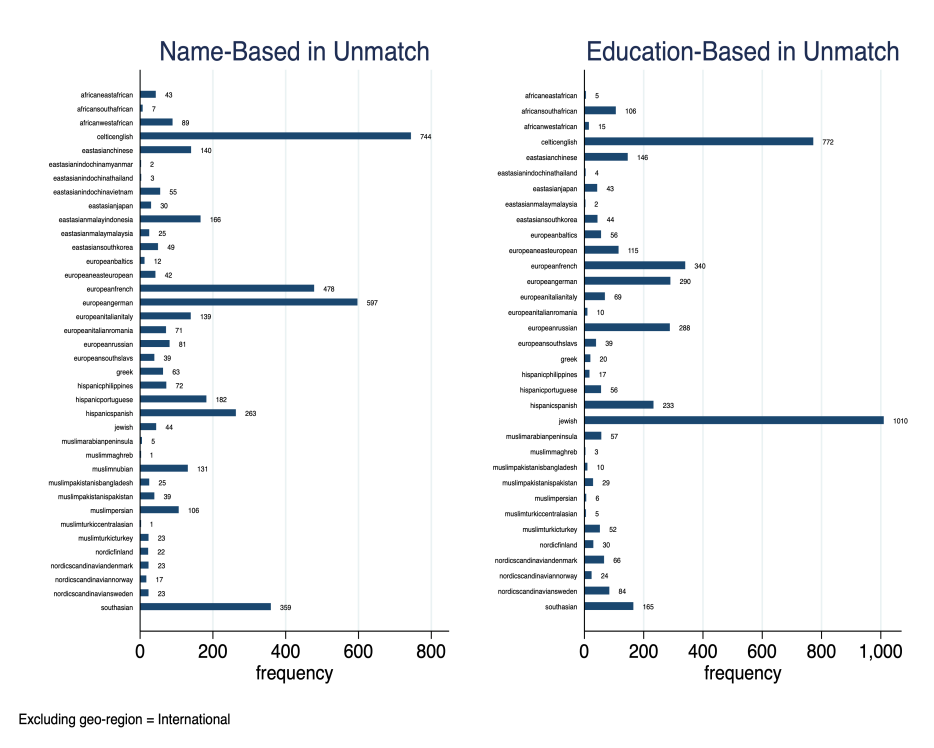

Figure 6: Distribution for Unmatched under Each Classification

In specific, Figure 8 and 9 demonstrate similar heat maps but conditional on the total mismatch of one category under one classification method. Mathematically, each point in Figure 8 and 9 shows the probability expressed by

$$\text{Prob}(MisMatch_{i_a,j_{a^c}}) = \frac{\#\ of\ MisMatch_{i_a,j_{a^c}}}{\sum_{j\neq i} \#\ of\ MisMatch_{i_a,j_{a^c}}}$$

where

- $i, j \in$ Nationality Taxonomy category set;
- $a, a^c \in$ {Name-based, Education-based} (if a is Name-based, then a^c is Education-based;
- $MisMatch_{i_a,j_{a^c}}$ denotes a mismatch that is classified as category i under method a but category j under a^c;
- $\#\ of\ MisMatch_{i_a,j_{a^c}}$ is the number of observations which is classified as category i under method a but category j under a^c;

Each figure in 8 and 9 is to be read as follows:

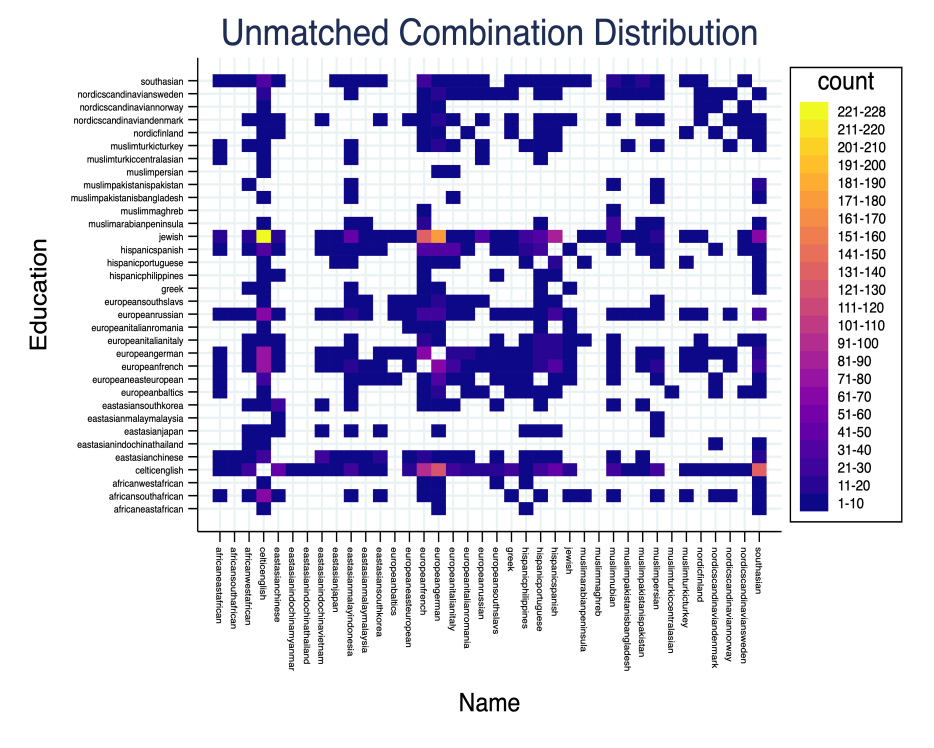

Figure 7: Distribution of Unmatched Combination

- Starts with Y-axis, each horizontal represents a category that observation is classified under the Y-axis-labelled method
- The rectangles on the horizontal line represent the probability that this specific category, classified under the Y-axis-labelled method, is classified as another category by the X-axis-labelled method.
- The brightness of each rectangle corresponds to the probability: the brighter the rectangle, the higher the probability is.

3.3 Sensitivity Analysis

This section shows the sensitivity of the matching percentage on the NamePrism probability threshold. Namely, the percentage of Education-Name match conditional on NamePrism probability is greater or equal to the probability threshold (excluding the "International" category from Education-based classification). Here are the results for three threshold benchmarks:

- When Prob($Match$) ≥ 0.25, the percentage of match is 57.67%;
- when Prob($Match$) ≥ 0.5, the percentage of match is 66.55%;
- when Prob($Match$) ≥ 0.75, the percentage of match is 73.17%;

Figure 10 shows the trend of match percentage over NamePrism matching probability thresholds. It is, as anticipated, upward-sloping against the threshold.

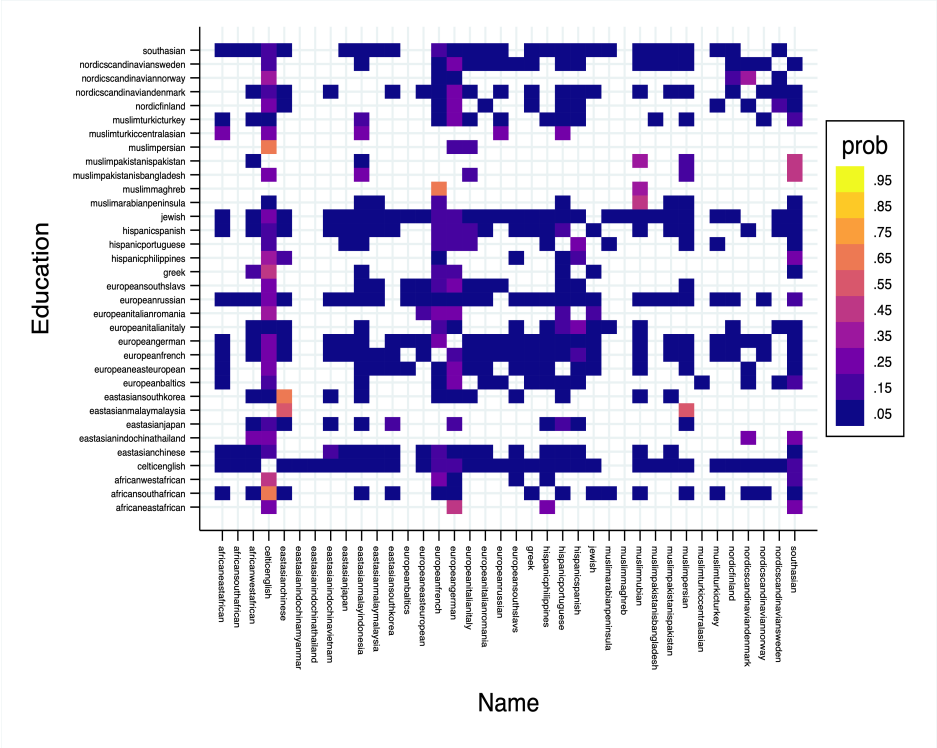

Figure 8: Distribution of Unmatched Combination Cond. on Education Classification

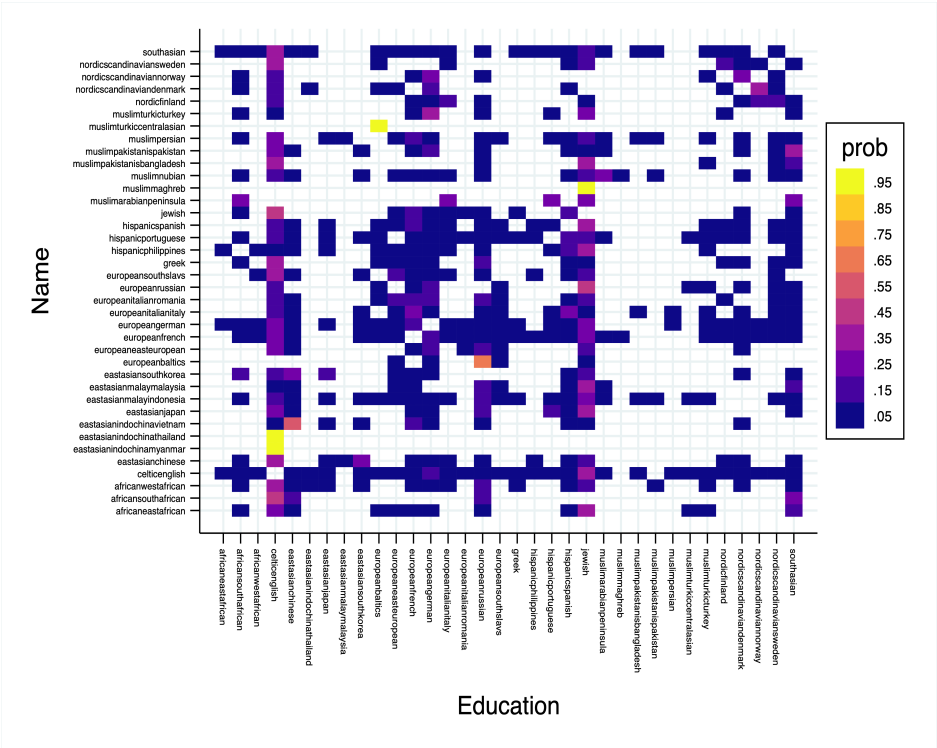

Figure 9: Distribution of Unmatched Combination Cond. on Last Name Classification

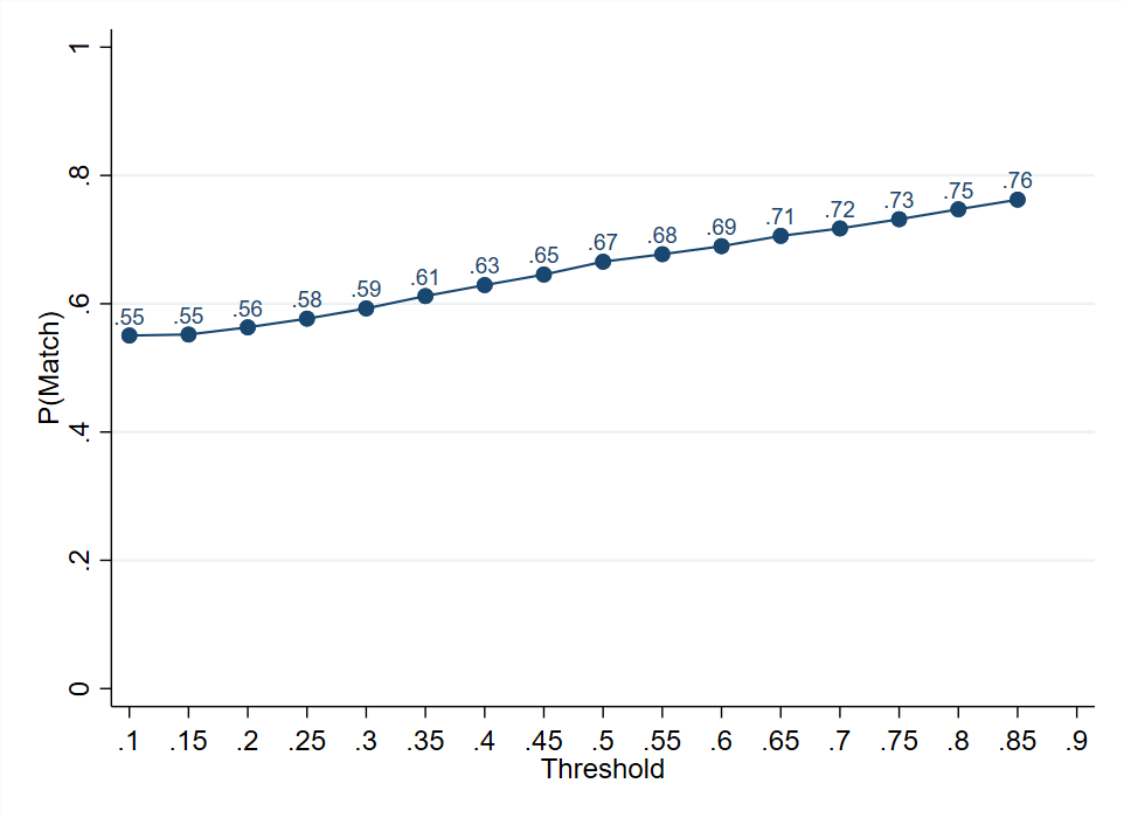

Figure 10: Match Percentage over NamePrism Probability